Open source media mining platform

Gérard Dupont
Research engineer – COEDS2 – Advanced studies
Media mining platform

From unstructured data from any sources...

... to structured and actionable knowledge
OSINT challenges

Some activities need to be automated:

- Search/Sources assessment
- Data Acquisition
- Classification, Screening, Indexing
- Information retrieval
- Knowledge capitalisation
- Visualization
- Summary
- Alert

Some activities cannot be automated:
- experts analysis of content;
- linking and mapping heterogenous information;
- evaluating reliability and assessing information;
- report and synthesis of information.

→ Tools can provide support but keep human in the loop.
An international Greenpeace alpine team delivers messages of support and hope for the victims of the nuclear disaster at Fukushima Daiichi to the summit of Mt Fuji. Collected from thousands of people in Japan and all over the world, Greenpeace hopes that these messages will help unite the people of Japan in opposition to nuclear power.

国際グリーンピース高山チームは富士山の頂上への支援と福島第一に原子力災害の被害者のための希望のメッセージを配信します。日本と世界中の何千人もの人々から収集した、グリーンピースは、これらのメッセージは、原子力発電に反対する日本の人々を締め込む役立つことを期待しています。
Integration approach

A platform providing "plug & play" functionalities for the integration of tools for collection, processing, analysis and communication...
Technology pile

Java Application Server

Enterprise Service Bus

Content store

Database

Portal

Maps server

Apache Tomcat

petals

Apache Jackrabbit

PostgreSQL

Liferay

GeoServer

SOA

ESB JBI

Portail/Portlets JSR168

BPEL

OWL

WSDL

SPARQL

RDFS

SOAP

XSD

XPath

XQuery

RDF

XML

Namespaces

URI

UTF-8

Apache Tomcat

Java Application Server

Enterprise Service Bus

Content store

Database

Portal

Maps server

OW2Con'12, November 28-29, 2012
Standard model & interfaces

WebLab Model 1.2.5

- «ComplexType» Annotation
  - PieceOfKnowledge
  - data : anyType
  -annotation [0..*]
  -pok [0..1]

- «ComplexType» ComposedResource
  - resource [0..*]

- «ComplexType» Resource
  - uri : anyURI

- «ComplexType» Query
  - operator
  - operator
  - weight : float
  - request : string

- «ComplexType» ComposedQuery
  - «ComplexType» ElementaryQuery
    - «ComplexType» Text
      - content : string [0..1]
    - «ComplexType» StringQuery
      - request : string
    - «ComplexType» SimilarityQuery

- «ComplexType» MediaUnit
  - mediaUnit

- «ComplexType» Segment
  - segment [0..*]
  - coordinate [2..*]
    - x : int
    - y : int

- «ComplexType» TemporalSegment
  - start : int
  - end : int

- «ComplexType» SpatialSegment
  - start : int
  - end : int

- «ComplexType» TrackSegment
  - data : anyType

- «ComplexType» LinearSegment
  - start : int
  - end : int

- «ComplexType» SpatioTemporalSegment
  - timestamp : int

- «ComplexType» Document
  - «ComplexType» Audio
    - content : base64Binary [0..1]
  - «ComplexType» Image
    - content : base64Binary [0..1]
Standard model & interfaces

Service

- Backbone_services
  - Development_factory
  - Monitoring
  - Orchestrator
  - Repository
    - Content_repository
    - Knowledge_repository_service
    - Resource_container_service
  - Service_bus
  - Web_portal

- Gathering_services
  - FocusGathering
  - StreamGathering

Processing_services

- Categorisation
  - Audio_Categorisation
  - ImageCategorisation
  - Text_Categorisation
    - Language_Identification
    - Topic_detection
  - Video_Categorisation
  - Cleansing
  - Information_Extraction
  - Media_Conversion
    - Audio_Extraction
    - Key_Frame_Extraction
    - Optical_character_recognition
    - Speech_Synthesis
    - Speech_Transcription
  - Media_Splitting
  - Near-Duplicate_Detection
  - Normalisation
  - Semantic_Enrichment
  - Summarisation
  - Translation

Exploitation_services

- Clustering
- Indexing
- Alert
- Information_Retrieval
  - Content_based_search
  - Full_Text_Search
  - Fuse_Search
  - Geographical_Search
  - Semantic_Search
  - Search_Support
    - Interactive_Search_Support
    - Result_Set_Enrichment
Application per domain
Application per domain
Application per domain
WebLab: a mature project

Applications dedicated to specific domains
- Business intelligence
- Knowledge management
- Military intelligence
- Medias Analysis

Integration of unstructured information processing softwares
- COTS
- Open source
- Advanced research components

Open source SOA infrastructure for unstructured information processing components
- Enterprise Service Bus (ESB)
- Portal
- Orchestrator
- Exchange data model

Sample application
- Local data collection
- Simple information extraction
- Text index/search

24 components including
- 8 technical services
- 9 services
- 7 portlets

Core platform including
- Data model
- ESB
- Portal
- Orchestrator
WebLab: a mature project

Page statistics

<table>
<thead>
<tr>
<th>Content pages</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages (All pages in the wiki, including talk pages, redirects, etc.)</td>
<td>349</td>
</tr>
<tr>
<td>Uploaded files</td>
<td>76</td>
</tr>
</tbody>
</table>

Edit statistics

| Page edits since WebLab Wiki was set up | 2094 |
| Average edits per page | 6 |

User statistics

| Registered users | 18 |
| Active users (Users who have performed an action in the last 30 days) | 4 |

View statistics

| Views total | 309 339 |
| Views per edit | 147.73 |

Technical stuff:
- code (SVN);
- bug tracking (JIRA);
- daily build (BAMBOO);
- code quality (SONAR);
- mailing list (8 guys).

Available tools:
- Maven plugin;
- Eclipse wizard;
- SOAPui test library;
- CLI test tools;
- Complete Bundle;
- ...
Thanks for your attention

Take away [weblab.ow2.org]

Logos and names of the tools presented are the property of their respective providers and are here only as illustration purposes on the already integrated technology in WebLab. Neither CASSIDIAN, nor EADS, claims any paternity on these external tools.

HERITRIX - http://crawler.archive.org
FFMPEG - http://ffmpeg.org/
SPHINX - http://cmusphinx.sourceforge.net/sphinx4/

GOOGLE TRANSLATE - http://translate.google.com/
GATE - http://gate.ac.uk/
JENA - http://jena.apache.org/